

工作周报

03.05.2012 - 03.18.2012

马昱欣

本周小结

很惭愧上周忘了写周报。本周周报总结最近两周的内容。主要工作是对Epinion等数据集中用户行为特征进行研究。

从三月初开始基本上每天都会做工作日志，是从学长那里学来的，发现对记录进度和督促工作有帮助。从开学以来工作上一直没有章法，比较急躁，希望这种方式能有所帮助。

每日进展

03.05.2012

今天的主要任务是继续计算Epinion数据集中用户的Pearson相关度。早晨使用Matlab (Statistical Toolbox) 自带的corr过程 (允许使用pearson参数) 计算。不过该方法并不属于Matlab中并行工具箱的一部分，方法仍旧和之前一样是串行执行的，没办法算出结果。顺便研究了一下Octave (Matlab的开源实现)，发现需要将底层的BLAS和LAPACK编译为并行版本才能实现并行运行。

下午开始总结一周以来对SocialTrustVis的总体思路。现在看起来整个流程可以串起来了，不过可视化方面的细节仍需深入思考。

晚上主要与彭老师讨论了下午的总结。彭老师主要指出两点：1. 对用户cluster特征的分析需要深化到语义层面，即具体地指出哪种特征对应哪种行为类型，以及是否存在“attacker”；2. 对cluster、individual之间的交互方式需要具体考虑。晚上同时重新阅读了Massa的《Trust Metrics on Controversial User》。提出了在SocialTrustVis中加入local和global metrics的对比，以更加客观地描述某个行为cluster。

对Pearson相关系数的计算陈老师建议使用CUDA加速。晚上粗略找了下相关文章，刚好有人直接给出了计算Pearson相关系数的CUDA实现。

03.06.2012

今天的内容主要有小组组会和CUDA平台搭建两部分。前一天晚上和彭老师已经讨论过大部分问题了，所以组会比较简短，主要是CUDA实现的内

容。CUDA计算，主要阅读了前一天找到的《Compute Pairwise Mahhatan Distance and Pearson Correlation Coefficient of Data Points with GPU》，还有一篇基于这篇文章的方法做Hierarchical Clustering的文章，两篇都主要讲了使用CUDA对Pearson相关度进行并行计算。

CUDA平台搭建比较麻烦。由于ATI显卡的缘故，自己的电脑只能使用模拟方式运行CUDA程序，而CUDA从3.0版本开始放弃对模拟方式的支持。2.3版本对模拟方式支持较好，但是在Lion系统上没办法加载CUDA.kext系统模块，即使使用32位内核方式也一样。于是换成CUDA 3.0，在32位内核方式下加载成功。期间用半个小时看了下OpenCL（Mac系统原生支持），发现入门资料太少，于是暂时放弃。编译CUDA程序时发现所有关于OpenGL的实例均编译失败，原因是Lion下的OpenGL Framework仅有x86_64一种架构支持，无i386架构。还好不会影响通用计算部分。找人借了本CUDA教程，大概熟悉了下kernel函数什么的，至少能看懂论文里给出的代码。

03.07.2012

今天主要工作是实现论文中的代码的IO部分，以及大组组会报告。IO部分对矩阵使用稀疏表达方法，空间占用很小。组会报告主要是将开学以来的工作进行汇报，主要使用的是之前发给老师们的proposal中的内容。组会上伟锋和海东二位学长的节能计算项目是组会讨论热点。

03.09.2012

今天主要是将之前的CUDA Kernel和IO部分连接起来，以及在模拟方式下进行调试。之前将GPU里面的grid、block和thread三个概念搞混了，现在才知道应该是一个thread对应结果矩阵中的一个空格。还有发现论文里面讲的kernel函数中计算Pearson相关度的方法并不适用于当前数据集，文中的所有向量维数相等且没有无关项，所以计算很快，而当前数据集中只对两个向量中对应位置同时为有效的维度进行计算，所以重新改写了kernel函数。还有改动了IO部分里面的结果矩阵存储形式为三角化存储，可以节省一半空间以放下更高维度的矩阵。晚上模拟方式调试通过并跑了第一组相似度结果，数据为Epinion数据集中打分个数大于50个的用户。

03.10.2012 - 03.12.2012

三天时间主要在完成毕设开题报告，顺便将之前的系统思路具体化一下。Parallel Spectral Clustering的Matlab实现性能还好，将3500个用户（打分个数大于50的用户）聚为2 - 200类，每次聚类操作需要10 - 20秒左右。还有将Slashdot和Wikivote两个数据集处理成CUDA程序需要的格式以便下一步计算。

03.13.2012

今天的主要工作是重写了Pearson相似度的CUDA程序数据接口，以及重新计算Epinion数据集和Wikivote数据集。

数据方面加入了命令行参数，现在可以接收指定格式的数据，方便处理其它数据集的数据。

Epinion数据集之前没有计算打分个数超过20和9个的用户，现在可以计算9个（平均打分个数）以上的用户（13000个左右），聚类速度在30秒左右。

03.14.2012

今天主要读了两篇有关Slashdot、Wikivote和Epinion三个数据集的文章：《Predicting Positive and Negative Links in Online Social Networks》和《The Slashdot Zoo: Mining a Social Network with Negative Edges》，以及计算Wikivote数据集的相似度。

第一篇文章重点在于使用机器学习的方法进行用户间的打分预测，与其它文章不同的是它使用了负评价这一因素，并且对正负评价的预测效果都很好。其中还讲到了一些transfer learning相关的内容，即从一个数据集上训练出的参数也适用于另一个数据集。这篇文章对用户行为建模贡献较少，略读而已。

第二篇文章重点分析了Slashdot这一数据集，主要研究了数据集本身的各种图论和统计方面特性。最后给出了一种基于降维的预测打分的方法。前半部分对数据集的特征研究可以帮助我们了解Slashdot数据集的分布情况。

晚上计算了Wikivote数据集的相似度（共~6100的用户，数据量较小），并计算了聚类结果，发现很多对同一用户打分行为不同的用户也被聚在了一起。

03.15.2012

今天详细看了之前Epinion数据集和Wikivote数据集的聚类结果，发现效果并不理想，从相似度矩阵看，发现有很多+1值（即两个用户的打分行为完全一致），但是考虑到每对用户间打分重叠的个数不尽相同，于是考虑到用户间的重叠打分个数也应作为参数加入相似度计算中。查了相关文章，发现Netflix报告中提到了对Pearson相似度进行放缩的方法，《Modeling Relationships at Multiple Scales to Improve Accuracy of Large Recommender Systems》中详细讲了对每对相似度乘上一个与用户间重叠打分个数有关的放缩系数，使得重叠打分个数更多的用户放缩后的相似度更大。晚上将改动后的相似度计算程序在所有数据集上重新跑了一遍。

用户行为划分方面找到了一篇2008 ACM Workshop on SocialNet上的文章《Identifying User Behavior in Online Social Network》，其中使用了YouTube中用户的Subscription信息作为用户间的打分关系，并根据一定的用户特征进行聚类的方法，将用户分成“圈中人”、“上传者”、“消费者”等5类。我们可以试着这篇文章对我们研究的数据集中的用户行为设定。

03.16.2012

今天的内容主要是小组组会。会上和彭老师交流了用户行为方面的思路和修改后的聚类结果分析。还有发现了斯坦福的Xiaolin Shi团队在社交网络用户行为方面的研究，略读了主页上的相关文章《User Grouping Behavior in Online Forums》等。

下周计划

- (1) 对用户行为方面进行研究，明确出需要表达的用户行为
- (2) 继续改进聚类结果，加入其他特征的参数